Chapter 2

Communicating molecular characteristics of disease

— documentation of diagnostic signatures —

Synopsis: *In this chapter I present methodology for the unambiguous documentation of molecular signatures. After an introduction to the preprocessing of microarray data I develop the key point: The scale inferred by preprocessing study data has to be documented along with the signature. I alter two popular preprocessing schemes to implement such a scale-conserving documentation. Using eight cancer data sets I am able to show that this kind of documentation leads to significantly less ambiguity in subsequent diagnoses as compared to standard approaches.*

2.1 Motivation

Microarray based gene expression signatures have the potential to be powerful tools for patient stratification, diagnosis of disease, prognosis of survival, assessment of risk group and selection of treatment [112, 130, 152]. These signatures are computational rules for deriving a diagnosis from a patient's expression profile, as discussed in the previous chapter. Before gene expression signatures can impact clinical practice, they need to be communicated to other health care centers with data for external evaluation [130], and ultimately to practitioners for use in clinical routine. This requires unambiguous documentation of the signature.

It has been observed that the reconstruction of published signatures can require an expert re-analysis of the study [141] or may not be possible at all [140]. Documenting a signature is conceptually different from reporting a list of genes contributing to the classification rule; the latter does not determine how a patient should be diagnosed. Moreover, also the exact quantitative specification of the classification rule does not constitute a *ready-to-use* signature. In addition, a procedure transforming raw expression data to a signature-specific scale has to be provided.

To motivate such a scale adjusting transformation we sketch an analysis step characteristic for studies involving microarray data: *Preprocessing*. Section 2.2 starts with an introduction and continues with more details on the two popular preprocessing schemes `vsn` and `rma`. In Section 2.3 we discuss why preprocessing the data implies the need of a scale adjusting transformation prior to classifying previously unknown samples. We provide such transformations for the preprocessing schemes discussed beforehand. The results are applied to real life data in Section 2.4. We assess eight clinical microarray studies [10, 12, 14, 65, 107, 118, 127, 153] and find that documenting signatures using standard procedures leads to unstable diagnoses. An independently diagnosed patient might well receive a diagnosis different from the one he would have received in the original study. We are able to show that scale adjusting transformations greatly reduce such ambiguity of diagnoses. In Section 2.5 we explore the possibility to utilize the `vsn` scale adjusting transformation to assess patient compatibility to core data. We critically discuss our findings providing guidelines for signature documentation in Section 2.6.

2.2 Preprocessing of oligonucleotide microarrays

Raw microarray data is subject to noise. There is variation in the data that is not due to biological signal, but rather to measurement error or experimental artifacts. Prior to data analysis it is therefore common practice to perform data *preprocessing*. Preprocessing of oligonucleotide microarrays generically encompasses three steps, although some methods may summarize more than one step in a single procedure:

1. *Background correction* This step comprises the identification of and correction for random signals that are not associated with `mRNA` abundance (background). Examples include optical noise and unspecific (cross) hybridization as well as other random sources of variation.
2. *Normalization* The normalization procedure aims to make measurements on different arrays comparable, removing systematic biases. Biases might arise through variations in amplification efficiency, signal quantification or human influences.
3. *Probeset summary* For oligonucleotide arrays, the background corrected and normalized measurements for different probes have to be aggregated to yield estimates of expression for the target sequences (genes).

Different methods exist for performing each of the three steps, while other methods aggregate more than one step at a time. It is not our aim to give an overview of state of the art approaches. Preprocessing of oligonucleotide microarrays is an active research field, and readers interested in the topic are referred to recent literature [61, 75, 100, 154, 160] and references therein. We note, though, that it is common practice to adopt a "plug-in" approach. Methods are combined in various ways and subsequent analysis stays conceptually decoupled from preprocessing. This has been reported to be suboptimal, as subsequent results may depend on arbitrary choices regarding the preprocessing scheme [61]. Recent works [63, 154] introduce a more principled approach combining preprocessing and certain subsequent analysis tasks in a model based approach. In the following, we briefly review two popular preprocessing schemes, for which we will derive a scale adjusting transformation later in Section 2.3.

2.2.1 Variance stabilization (`vsn`)

In this section we discuss a preprocessing strategy utilizing a variance stabilizing transformation to normalize microarray data, originally introduced by Huber et al. [67, 68]. We refer to this preprocessing scheme as the `vsn` scheme. It combines background correction and normalization by specifying a stochastic model for gene expression values. Methodology for robust estimation of the model parameters is implemented in the Bioconductor [54] package `vsn`.

The `vsn` model

Huber et al. [67] specify a model for the measured `mRNA` abundance x_{ki} of probe k in sample i:

$$\begin{aligned} x_{ki} &= \alpha_{ki} &&+ \beta_{ki} y_{ki} \\ &= (a_i + \nu_{ki}) &&+ (\beta_i \gamma_k e^{\eta_{ki}}) y_{ki} \quad , \end{aligned}$$

where y_{ki} represents the true abundance. The measured abundance has been decomposed into an unspecific signal contribution α_{ki} and a term $\beta_{ki} y_{ki}$ linearly dependent on the true abundance. Further on, the unspecific term has been split up into a per-sample offset a_i and an additive noise component ν_{ki} that does not depend on the amount of `mRNA` present. The proportionality factor β_{ki} has been further split up into a probe-specific affinity γ_k and a sample specific overall normalization factor β_i. A multiplicative noise component $e^{\eta_{ki}}$ has been introduced as well. To make the model identifiable the constraints $\sum_{k=1}^{p} \nu_{ki} = 0$, $\sum_{i=1}^{n} \eta_{ki} = 0$ and $\sum_{k=1}^{p} \eta_{ki} = 0$ are imposed on the noise terms. This fixes the scales of β_i and γ_k and it enables the interpretation of a_i as a chip-specific signal offset.

Expressing the `mRNA` abundance in probe specific units $m_{ki} = \gamma_k y_{ki}$ and taking the ν_{ki} and η_{ki} to be realizations of zero mean random variables implies the following model:

$$\frac{X_{ki} - a_i}{\beta_i} = m_{ki} e^{\eta_{ki}} + \frac{\nu_{ki}}{\beta_i} \quad .$$

Assuming the random variables ν_{ki} to be iid with variance σ_ν^2 and the η_{ki} to be iid with variance σ_η^2 the mean and variance of X_{ki} turn out to be:

$$\begin{aligned}\mathbb{E}[X_{ki}] &= \beta_i m_{ki}\mathbb{E}[e^\eta] + a_i\\ \mathrm{Var}[X_{ki}] &= (\beta_i m_{ki})^2\mathrm{Var}[e^\eta] + \beta_i^2\sigma_\nu^2\\ &= (\mathbb{E}[Y_{ki}] - a_i)^2\frac{\mathrm{Var}[e^\eta]}{\mathbb{E}[e^\eta]^2} + \beta_i^2\sigma_\nu^2 \quad .\end{aligned} \tag{2.1}$$

We have dropped the subscripts of the random variables ν and η, as they are same for all probes k on all arrays i. Correspondingly, we are not interested in the effect of σ_η on the mean of X_{ki} and regard it as constant. Then Equation (2.1) describes a variance to mean relation with a constant and with a quadratic contribution. This implies two regimes. For highly expressed genes the quadratic term dominates and implies enlarged variance. For weakly expressed genes with $\mathbb{E}[X_{ki}]$ close to a_i, the constant term $\beta_i^2\sigma_\nu^2$ dominates. Its contribution can be viewed as the background noise level of sample i, which is present regardless of the magnitude of gene expression. These two regimes are a direct consequence of the combination of an additive error term and independent multiplicative one in the underlying model. This type of models has also been proposed by Durbin et al. [38] and was more recently taken up by Wu et al. [155], as the quadratic mean to variance relation agrees well with empirical evidence. We note, though, that a slightly different variance to mean dependence (e.g. Kepeler et al. [82] propose a power law with an exponent less than two) cannot be ruled out.

The variance stabilizing transformation

Analysis of the gene expression estimates can benefit from a constant variance function, and (approximately) variance stabilizing transformations have been proposed in [39, 67, 71, 116, 117], among others. The idea starts with viewing expression estimates as realizations of random variables with a mean depending on the magnitude of the (true) expression. Then a smooth transformation h is derived, such that the transformed random variable's variance is independent of its mean. If the same transformation is applied to the expression estimates, one expects the sample variance to be independent of the sample mean. The transformation h can be found by Taylor expansion:

Let X_u be a random variable with mean u, and h a smooth function. Then

$$\begin{aligned}\mathrm{Var}[h(X_u)] &\approx \mathrm{Var}\big[h(u) + h'(u)(X_u - u)\big]\\ &= h'(u)^2\underbrace{\mathrm{Var}[X_u]}_{:=v(u)} \quad .\end{aligned} \tag{2.2}$$

Setting the transformed variance to one yields h via

$$h'(u) = \frac{1}{\sqrt{v(u)}} \quad \text{or}$$
$$h(y) = \int^{y} 1/\sqrt{v(u)}du \quad .$$

With a variance to mean dependence as in Equation (2.1) the integrand is of the type $1/\sqrt{u^2+\alpha}$ and the integral is known to be the *asinh* function:

$$h(X_{ki}) = \operatorname{asinh}\frac{X_{ki}-a_i}{b_i}$$

with $b_i = \beta_i\sigma_v/c$ and $c = \sqrt{\mathrm{Var}[e^\eta]/\mathbb{E}[e^\eta]^2}$. Correspondingly, the variance stabilized model now reads:

$$\operatorname{asinh}\frac{X_{ki}-a_i}{b_i} = \mu_{ki} + \varepsilon_{ki} \quad , \tag{2.3}$$

where $\mu_{ki} = \mathbb{E}[\operatorname{asinh}\frac{X_{ki}-a_i}{b_i}]$ is the (transformed) expression estimate for gene k in sample i, and ε_{ki} are iid random variables with zero mean and constant variance $\sigma_\varepsilon^2 = c^2$. While the quality of the approximation of h (in Equation (2.2)) using Taylor expansion depends on the concentration of X_u around u, empirical evidence suggests that the method works well for microarray data.

Background correction and normalization

Huber et al. [67] provide an estimation procedure for the parameters $\{a_i, b_i\}_{i=1}^n$ in Equation (2.3), based on the assumption that the majority of probes are equally expressed in all samples. It is based on the model of Equation (2.3):

$$\operatorname{asinh}\frac{X_{ki}-a_i}{b_i} = \mu_k + \varepsilon_{ki} \quad , \quad \varepsilon_{ki} \sim N(0,\sigma_\varepsilon^2)$$

and parameter estimates $\{\hat{a}_i, \hat{b}_i\}_{i=1}^n$ are obtained by the maximum likelihood method. The additional unknowns μ_k and σ_ε^2 are treated in terms of a *profile likelihood* [101]. Robustness against the normal assumption is achieved in the tails by using a *least trimmed sum of squares* [120, 121] approach, which also yields a set of approximately constant genes $\mathcal{K}$. The background corrected and normalized expression estimate of probe k in sample i is given by:

$$x_{ki}^{norm} = \hat{h}_i(x_{ki}^{\mathtt{raw}}) = \operatorname{asinh}\big((x_{ki}^{\mathtt{raw}} - \hat{a}_i)/\hat{b}_i\big) \quad .$$

Probeset summary

Let us collect the background corrected and normalized probe level expression estimates in a $p \times n$ matrix $\boldsymbol{X}$. Different methods have been proposed to summarize the probe level data to an expression measure per gene. There seems to be some evidence favoring model based approaches [72, 88] over more heuristic concepts [73, 160]. We use an additive model [72] with the following rationale: Each probe binds labelled target fragments with a certain efficiency. This is encoded in the `vsn` model, as we are estimating the probe expression in probe specific units. These units can be assumed to be the same across different arrays, but different for the various probes contributing to the same probeset. When aggregating information of the different probes, this can be taken into account by estimating these scales from the data at hand.

Denote by $\boldsymbol{X}^{(k)}$ the submatrix of the probes (across all arrays) assigned to probeset targeting sequence k. Then an additive model assumes

$$\boldsymbol{X}^{(k)}_{ij} = p_{ki} + g_{kj} + \epsilon \quad ,$$

where p_{ki} is a probe specific effect encoding the probe specific units while g_{kj} represents the abundance of `mRNA` of gene k in array j. That the scale p_{ki} is estimated by an *additive* constant and not multiplicatively, arises from the fact that we have transformed the data with the *asinh* function. Looking at the two regimes of highly and weakly expressed genes we find:

$$\begin{aligned} \operatorname{asinh}(x) &\approx \log(2) + \log(x) && \text{for} \quad x > 1 \\ \operatorname{asinh}(x) &\approx \operatorname{sgn}(x) \cdot \log(|x| + 1) && \text{for} \quad |x| < 1 \quad . \end{aligned} \tag{2.4}$$

We show the above approximation in Figure 2.1. Since $\log(ax) = \log(a) + \log(x)$, the additive model for the scale comes naturally for well expressed genes. For weakly expressed genes a fudge factor is effectively included, reducing the effect of inflated fold changes in this regime. Estimates for the probe scales $\hat{p}_{ki}$ and expression values $\hat{g}_{kj}$ are obtained via the median polish procedure [144]. Row and column medians are iteratively swept out of $\boldsymbol{X}^{(k)}$ until convergence; p_{ki} is estimated by the cumulative row medians, g_{kj} by the column medians. The $\hat{g}_{kj}$ then make up the gene level `vsn` preprocessed expression matrix.

2.2.2 Robust multichip average (`rma`)

The `rma` preprocessing protocol was proposed by Irizarry et al. [72]. It stands for a specific combination of methods for each of the three preprocessing steps of background correction, normalization and probeset summary. We present enough detail to be able to derive a scale adjusting transformation. Further information is available via the original publication [72] and in the documentation of the Bioconductor [54] package `affy`.

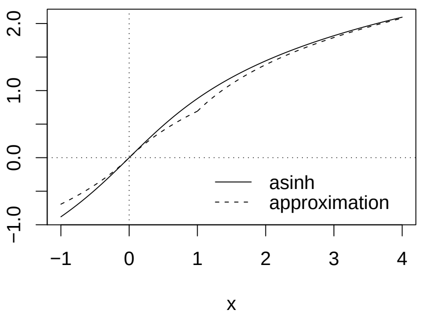

***Figure 2.1: Approximation of the* asinh *function**. We see that for large values the* asinh *function can well be approximated by the logarithm. For smaller values a constant has to be added (see Equation (2.4)). This motivates the use of an additive model for probeset summary.*

Background correction

For the background correction step the `rma` method uses the following model [72]. The measurements $x_{ki}^{\mathtt{raw}}$ (probe k, sample i) are assumed to consist of a signal (s_{ki}) as well as of a noise contribution (b_{ki}):

$$x_{ki}^{\mathtt{raw}} = b_{ki} + s_{ki} \quad .$$

The components b_{ki} and s_{ki} are treated as iid instances of two independent random variables. For each array the background distribution is assumed to be normally distributed ($b_{ki} \sim N(\mu_i, \sigma_i)$), while the signal contribution is assumed to come from an exponential distribution ($s_{ki} \sim Exp(\lambda_i)$). To avoid the possibility of negative values, the background distribution is truncated at zero. This yields for the expected signal, conditional on the observed measurement [52, 72]:

$$\mathbb{E}[s_{ki}|x_{ki}^{\mathtt{raw}}] = a_{ki} + \sigma_i \Big(\phi(\frac{a_{ki}}{\sigma_i}) - \phi(\frac{x_{ki}^{\mathtt{raw}} - a_{ki}}{\sigma_i})\Big) \;/\; \Big(\Phi(\frac{a_{ki}}{\sigma_i}) + \Phi(\frac{x_{ki}^{\mathtt{raw}} - a_{ki}}{\sigma_i}) - 1\Big)$$

with $a_{ki} = x_{ki}^{\mathtt{raw}} - \mu_i - \sigma_i^2 \lambda_i$, and ϕ, Φ are the density and distribution function of the normal distribution, respectively. Estimates $\hat{\mu}_i$, $\hat{\sigma}_i$ and $\hat{\lambda}_i$ are obtained from a per-chip kernel density estimate of the raw expression values. $\mathbb{E}[s_{ki}|x_{ki}^{\mathtt{raw}}]$ is used as a background corrected estimate of the expression values at probe level.

This background model was originally published with the `rma` method [72], and we choose to work with it because its application is frequently encountered in practice. More recent work on background correction / signal identification can be found for instance in [155], where probe composition is taken into account.

Normalization

For normalization `rma` employs *quantile normalization* [16]. The procedure is not model based, and the idea is the following. The distribution function of expression measurements is assumed to be the same for each array. On a population scale, the expression values do not change from sample to sample. This is not encountered in practice, and as a correction to systematic biases the empirical distribution functions of expression values on each array are *made* to be the same.

In more detail it works like this: Denote by $\boldsymbol{X}$ the $p \times n$ `rma` background corrected probe-level expression matrix on log scale. That is, $\boldsymbol{X} = \left\{\log(\mathbb{E}[s_{ki}|x_{ki}^{\texttt{raw}}])\right\}$. Let Π be the permutation sorting the columns of $\boldsymbol{X}$ and Π^{-1} its inverse. Then the quantile normalized version of $\boldsymbol{X}$ is obtained via:

$$\tilde{\boldsymbol{X}} = \Pi^{-1}((\Pi\boldsymbol{X})\mathbf{1}) \quad ,$$

where $\mathbf{1}$ is a $n \times p$ matrix with all elements equal to $1/n$. The empirical distribution function on all arrays is then the same. All probes with the same rank have the same expression estimate on all arrays, namely the mean (across the arrays) of all estimates of probes with this rank.

Probeset summary

For probeset summary the same additive model as discussed in the `vsn` case is used. Expression estimates on gene level are then also provided by the $\hat{g}_{kj}$.

2.3 Documentation of signatures

After having introduced two preprocessing schemes, we investigate the effect of preprocessing on documentation requirements for diagnostic signatures. Such signatures are quantitative computational rules, deriving a diagnosis from a patient's expression profile. Generally, signatures are derived from example data by statistical learning techniques, as described in the first chapter. The point we make is the following: As signatures are inferred from preprocessed data, crucial information on the preprocessing is not encoded in a quantitative description alone. A properly documented signature provides a scale adjusting transformation for future data by aggregating preprocessing information.

We continue discussing how preprocessing implies the need for a scale adjusting transformation and derive two such transformations for the preprocessing schemes `vsn` and `rma` discussed in Section 2.2.

2.3.1 Preprocessing implies a scale adjusting transformation

Microarray data is not measured on an absolute scale; it is prone to heteroscedastic noise and systematic biases. Therefore it is common practice to preprocess the data prior to analyses such as signature derivation, as discussed in Section 2.2. The preprocessing step can be viewed as a procedure that finds a common scale and transforms all arrays in a data set. This enables meaningful comparison of expression estimates between the arrays. Generally, the preprocessed expression estimate x_{ki} of gene k in array i is inferred from the arrays A in a data set $\mathscr{D} = \{A_1, \dots, A_n\}$ with a preprocessing scheme, say prep. Typically preprocessing algorithms additionally depend on user adjustable parameters p. In summary we can write this as:

$$x_{ki} = x_{ki}^{\mathscr{D}} = \texttt{prep}(x_{ki}^{\texttt{raw}}) = \texttt{prep}(x_{ki}^{\texttt{raw}}; A_1, \dots, A_i, \dots, A_n, \mathrm{p}) \quad ,$$

where we should actually write $\{x_{li}^{\texttt{raw}}\}_{l \in K}$ with K the set of probes coding for target sequence k. As this is rather cumbersome we prefer to overload k, as the meaning is usually clear from the context.

Now consider a generic situation in molecular diagnostics: Derivation and application of a molecular signature. Given a core data set we hope to infer a signature, using a learning algorithm. We plan to apply the signature to other (new) data to determine disease type or predict disease outcome. The learning algorithm is applied to *preprocessed* core data to infer the signature. The resulting classifier is tuned to the scale of the input data; this scale, in turn, has been determined by the preprocessing scheme. To classify a new sample $A_{(n+1)}$, this sample has to be transformed to the same scale. To be applicable, the signature $c_{\mathscr{D}}$ therefore has to be accompanied by a transformation $f_{\texttt{prep},\mathscr{D}}$ such that $f_{\texttt{prep},\mathscr{D}}(A_{(n+1)})$ is comparable to the samples in $\mathscr{D}$; then we can classify the new patient via $c_{\mathscr{D}}(f_{\texttt{prep},\mathscr{D}}(A_{(n+1)}))$. This implies a properly documented signature consists not only of a quantitative description of the classification rule, but also includes a transformation enabling its straight forward application to new raw data.

Two kinds of preprocessing schemes

As mentioned, the scale adjusting transformation $f_{\texttt{prep},\mathscr{D}}$ depends on the preprocessing scheme used. Conceptually we distinguish two cases. If a preprocessing scheme does not share information across the arrays, the samples can be processed independent of each other:

$$x_{k(n+1)} = \texttt{prep}(x_{k(n+1)}^{\texttt{raw}}; A_1, \dots, A_{(n+1)}, \mathrm{p}) = \texttt{prep}(x_{k(n+1)}^{\texttt{raw}}; A_{(n+1)}, \mathrm{p}) \quad .$$

In that case, we can choose the scale adjusting transformation via$f_{\texttt{prep},\mathscr{D}} = \texttt{prep}$. We can preprocess a new array with exactly the same algorithm we used for the original arrays. Only the parameters p should be the same and need to be documented.

The second case comprises preprocessing schemes pooling information across arrays to estimate a common scale. For such procedures, the scale adjusting transformation depends on original data $\mathscr{D}$ as well on the preprocessing scheme. The reason is, that an expression estimate for a gene on one array depends on the raw data of the other arrays. Adding a new chip to the data set changes the expression estimate:

$$x_{ki}^{\mathscr{D}^+} := \mathtt{prep}(x_{ki}^{\mathtt{raw}}; A_1, \ldots, A_{(n+1)}, \mathtt{p}) \neq \mathtt{prep}(x_{ki}^{\mathtt{raw}}; A_1, \ldots, A_{(n)}, \mathtt{p}) =: x_{ki}^{\mathscr{D}} \quad , \tag{2.5}$$

where $\mathscr{D}^+ = \mathscr{D} \cup A_{(n+1)}$ and $i \leq n$. In case the original data is available, one may consider the approach $f_{\mathtt{prep},\mathscr{D}}(\cdot) = \mathtt{prep}(\cdot; A_1, \ldots, A_{(n+1)}, \mathtt{p})$ for the scale adjusting transformation. The assumption then is that adding a single array to $\mathscr{D}$ does not significantly alter the results of the preprocessing. The preprocessing algorithm is assumed to be robust with respect to changes to the input data, assuming near equality in Equation (2.5). This seems plausible for large data sets, but we are unaware of any systematic or empirical work on this topic. In the following we present a short analysis.

Perturbing the input data

Taking a data set on childhood Leukemia [153], we analyze empirically how addition of an extra array affects expression estimates of the "original" arrays. As preprocessing schemes we consider `rma` and `vsn`. To visualize this inclusion-effect, we plot histograms of $(x_{ki}^{\mathscr{D}^+} - x_{ki}^{\mathscr{D}}) / (\frac{1}{2}(x_{ki}^{\mathscr{D}^+} + x_{ki}^{\mathscr{D}}))$ for $i \leq n$.

We take $n = 20$ random arrays from the data set and exclude each chip once to form instances of $\mathscr{D}$ and $\mathscr{D}^+$; results are shown in Figure 2.2. The effect is about the same for both, `rma` and `vsn`, and not necessarily negligible. Further on, we examine the dependency of this inclusion-effect on the size of the original data set. In Figure 2.2 we took twenty arrays as the base-set. In Figure 2.3 we start with three samples as the base set and increase the size to 41 samples. As a measure of the overall effect for each base-set size (a summary of the histogram in Figure 2.2) we choose the inter quartile range (IQR). In Figure 2.3 we plot the IQRs in dependence of the base-set size. The solid lines correspond to a fit of $f(n) = \alpha / n$, which seems to describe the size dependency adequately. We hypothesize that choosing $f_{\mathtt{prep},\mathscr{D}}(\cdot)$ to be $\mathtt{prep}(\cdot; A_1, \ldots, A_{(n+1)}, \mathtt{p})$ is reasonable for large data sets, but not optimal. Also, empirically looking at one data set is clearly not enough evidence for a general rule.

To recapitulate, we have discussed why preprocessing implies the need of a scale adjusting transformation. We briefly assessed the effect of re-processing the concatenated data set as a substitute for a scale adjusting transformation. In the following we give advantages of a direct approach to $f_{\mathtt{prep},\mathscr{D}}$:

- A direct approach addresses the correct question. In the scenario described, we are not interested in estimating a *new* scale for more arrays, but to transform an array to an *already existing* scale.
- Arrays can be added to existing data sets without affecting expression estimates. This ensures consistency between preliminary and final analyses in studies where data keeps being added.

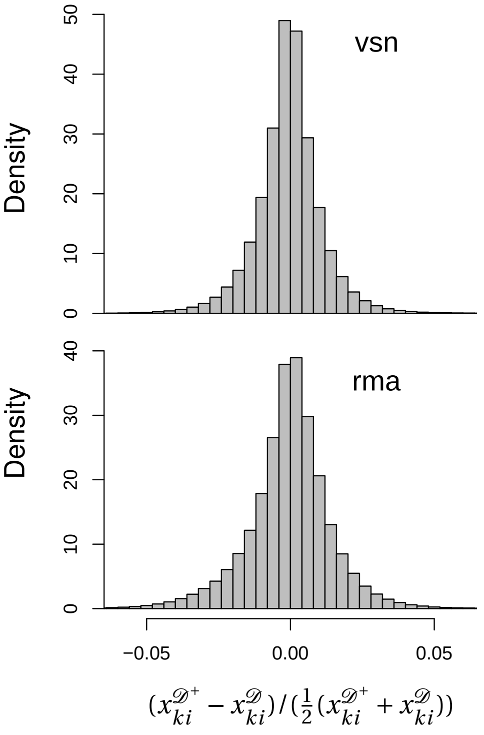

Figure 2.2: Effects of inter-array dependencies in two preprocessing schemes. *For the two preprocessing schemes* `vsn` *and* `rma` *we show a histogram of the* $(x_{ki}^{\mathscr{D}^+} - x_{ki}^{\mathscr{D}})/(\frac{1}{2}(x_{ki}^{\mathscr{D}^+} + x_{ki}^{\mathscr{D}}))$ *(see text). This demonstrates how the addition of an extra array to a data set affects the expression estimates of the original arrays via preprocessing (see text). We take twenty random arrays from a microarray data set [153] and exclude each chip once to form an instance of* $\mathscr{D}$ *and* $\mathscr{D}^+$*. The effect is about the same for both schemes, and this experiment shows that effects of each single array on all others of a data set are not necessarily negligible.*

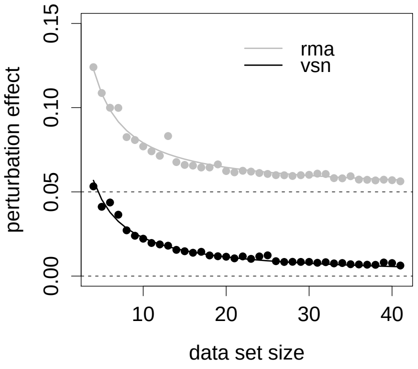

Figure 2.3: Size dependency of the effects of inter-array dependencies for two preprocessing schemes. *We plot a perturbation measure (see text) over the size of the data set* $\mathscr{D}^+$*. The* `rma` *plot is shifted vertically (dashed lines). We see that the perturbation is generally small (about 5%) and roughly proportional to the inverse of the data set size (solid lines).*

- While the computational benefits in detail depend on the preprocessing schemes, our experience suggests they can be substantial in terms of memory requirements as well as in computation time.
- Defining a scale by the entire raw study data and a reference to a preprocessing protocol is clearly suboptimal. Providing an independent $f_{\mathtt{prep},\mathscr{D}}$ allows an easy and meaningful exchange of signatures between scientists. This is beneficial to speed up the independent evaluation of molecular signatures, a crucial step on their way to clinical practice.

In the following we provide $f_{\mathtt{prep},\mathscr{D}}$ directly for the `rma` and `vsn` preprocessing schemes.

2.3.2 A scale adjusting transformation for `vsn` preprocessed signatures

In Section 2.2 we described the preprocessing procedure `vsn`, based on a variance stabilizing transformation [67] in conjunction with an additive-multiplicative error model. In the following we derive the associated scale adjusting transformation $f_{\mathtt{vsn},\mathscr{D}}$.

Recall, that the `vsn` procedure consists of two sequential steps. The inference of the variance stabilizing transformation (normalization on probe level) is followed by the probeset summary. Accordingly, $f_{\mathtt{vsn},\mathscr{D}}$ will be a two step procedure: $f_{\mathtt{vsn},\mathscr{D}} = f^{sum}_{\mathtt{vsn},\mathscr{D}} \circ f^{norm}_{\mathtt{vsn},\mathscr{D}}$. The first step will be analogous to the inference of the variance stabilizing transformation, the second to the probeset summary.

Deriving the normalizing transformation

In the `vsn` preprocessing scheme a normalizing transformation for all the arrays in a data set is inferred from a stochastic model (Equation (2.3)). The underlying assumptions are that for the majority of genes expression remains unchanged across arrays; and that the constant genes, after a variance stabilizing transformation, scatter around some common mean value μ_k:

$$x_{ki} = \hat{h}_i(x^{\mathtt{raw}}_{ki}) = \operatorname{asinh}\big((x^{\mathtt{raw}}_{ki} - \hat{a}_i)/\hat{b}_i\big) = \mu_k + \epsilon_{ki}, \quad \epsilon_{k_i} \sim N(0, \sigma^2_\epsilon) \tag{2.6}$$

Equation (2.6) describes the expression estimate of probe k on array i. Heteroscedasticity of the raw data has been absorbed into the *asinh* transformation. The estimates $\{\hat{a}_i, \hat{b}_i\}^n_{i=1}$ are inferred by a maximum likelihood approach. The fitting procedure employed is robust against deviations from the normal assumption in the tails of the distribution governing the residuals ϵ_{ki}.

To derive the normalization step of $f_{\mathtt{vsn},\mathscr{D}}$, we start with a `vsn` normalized data set comprising n patients. From the procedure of Huber et al. [68] we are not only equipped with parameter estimates, but also with a set of nearly constant genes $\mathscr{K}$. Further on, estimates of the means ($\hat{\mu}_k$) and the variance ($\hat{\sigma}^2_\varepsilon$) in Equation (2.6) are also at hand:

$$\hat{\mu}_k \quad = \quad \frac{1}{n}\sum_{i=1}^{n} x_{ki} \quad \text{and} \quad \hat{\sigma}^2_\varepsilon \quad = \quad \frac{1}{n|\mathscr{K}|}\sum_{k\in\mathscr{K}}\sum_{i=1}^{n}(x_{ki}-\hat{\mu}_k)^2 \quad .$$

These estimates are used internally in the `vsn` fitting routine to calculate a profile likelihood. We utilize them to rewrite Equation (2.6) into a stochastic model for the expression values of an additional (external) patient:

$$x_{k(n+1)} = h_{n+1}(x^{\texttt{raw}}_{k(n+1)}) = \hat{\mu}_k + \epsilon_k, \quad \epsilon_k \sim N(0,\hat{\sigma}^2_\varepsilon) \text{ for } k \in \mathscr{K}. \tag{2.7}$$

The likelihood corresponding to the model is of the form

$$\mathscr{L}(A_{n+1}|a_{n+1},b_{n+1}) = \sum_{k\in\mathscr{K}} \frac{(h(x^{\texttt{raw}}_{k(n+1)})-\hat{\mu}_k)^2}{2\hat{\sigma}^2_\varepsilon} - \sum_{k\in\mathscr{K}} \log(\partial_x h(x^{\texttt{raw}}_{k(n+1)})) \quad , \tag{2.8}$$

where the dependence on the parameters a_{n+1} and b_{n+1} is through h. Maximum likelihood estimates $(\hat{a}_{n+1},\hat{b}_{n+1}) = \text{argmax}_{(a,b)}\mathscr{L}(A_{n+1}|a,b)$ can be obtained numerically and define a normalizing transformation for the $(n+1)$-th array:

$$f^{norm}_{\texttt{vsn},\mathscr{D}}(x^{\texttt{raw}}_{k(n+1)}) = \text{asinh}\big((x^{\texttt{raw}}_{k(n+1)} - \hat{a}_{n+1})/\hat{b}_{n+1}\big) \quad .$$

To be able to derive this normalizing transformation we only need information to calculate $\hat{a}_{n+1}$ and $\hat{b}_{n+1}$. That is, for a given data set $\mathscr{D}$, the parameters $\mathscr{K}$, $\hat{c}^2$ and the $\hat{\mu}_k$ define the scale adjusting transformation and should be documented.

Defining the probeset summary

As discussed in Section 2.2, the `vsn` preprocessing scheme uses an additive model to summarize probe level data. If $\boldsymbol{X}^{(k)}_{ij}$ denotes $f^{norm}_{\texttt{vsn},\mathscr{D}}(x^{\texttt{raw}}_{ij})$, the normalized expression estimate of the i-th probe of probeset k on the j-th array, this model assumes (see Section 2.2):

$$\boldsymbol{X}^{(k)}_{ij} \approx p_{ki} + g_{kj} \quad .$$

The probe effects p_{ki} and the array effects g_{kj} are estimated by the median polish procedure and are available from the original data. The array effects are the reported estimates of the expression value of the target sequence k, while the probe effects $\hat{p}_{ki}$ represent a probe-dependent scale. In compliance with the original additive model, we calculate the expression value of the k-th gene on the $(n+1)$-th array as

$$x_{k(n+1)} = \mathtt{median}\left(f^{norm}_{\mathtt{vsn},\mathscr{D}}(x^{\mathtt{raw}}_{i(n+1)}) - \hat{p}_{ki}\right) \quad ,$$

where the median is taken over all probes i belonging to probeset k. With this, we have available the probeset summary part of $f_{\mathtt{vsn},\mathscr{D}}$. To define this part of the scale adjusting transformation, we need the probe effects $\hat{p}_{ki}$ estimated from the original data set.

In summary, for the vsn preprocessing scheme a scale adjusting transformation $f_{\mathtt{vsn},\mathscr{D}}$ can be defined. It is dependent on aggregated information of the original data set: A set of not differentially expressed probes ($\mathscr{K}$) and means $\hat{\mu}_k$ around which they fluctuate with variance $\hat{\sigma}^2_\varepsilon$. For the probeset summary we need probe specific scales $\hat{p}_{ik}$ to weigh each probe's contribution to the expression estimate.

2.3.3 A scale adjusting transformation for rma preprocessed signatures

In the following we present a scale adjusting transformation for rma preprocessed data. For more details of the preprocessing algorithms see Section 2.2. As the rma preprocessing procedure is a three step algorithm, the scale adjusting transformation $f_{\mathtt{rma},\mathscr{D}}$ will be of the form $f_{\mathtt{rma},\mathscr{D}} = f^{sum}_{\mathtt{rma},\mathscr{D}} \circ f^{norm}_{\mathtt{rma},\mathscr{D}} \circ f^{back}_{\mathtt{rma},\mathscr{D}}$. Since the background correction step employed by rma considers arrays independently, we could employ the same background correction as the original algorithm. In fact, we will see that we do not need any background correction at all. As a summary method rma employs the same additive model we discussed before. Consequently, $f^{sum}_{\mathtt{rma},\mathscr{D}}$ is the same as for the vsn case. This leaves the normalization part to be defined.

Deriving the normalizing transformation

The rma procedure utilizes a method called quantile normalization [16]. The normalization procedure ranks all the genes on each array. Then it assigns each gene of a certain rank the mean expression value of all genes of this rank (see Section 2.2). Let $\hat{\boldsymbol{\mu}}$ be a vector collecting all those mean values and let Π denote the permutation sorting the background corrected expression estimates on the $(n+1)$-th array. Then quantile normalized expression values can be obtained via

$$f^{norm}_{\mathtt{rma},\mathscr{D}} \circ f^{back}_{\mathtt{rma},\mathscr{D}}(x^{\mathtt{raw}}_{k(n+1)}) \;=\; (\Pi^{-1}\hat{\boldsymbol{\mu}})_k \quad , \tag{2.9}$$

where Π^{-1} is the inverse of Π. Note though, that we do not need the background corrected (unnormalized) expression values to get hold of the permutation Π^{-1}. Since the rma background correction is a global (strictly) monotonous transformation, we can derive Π^{-1} directly from the raw data. Therefore the scale adjusting transformation for rma preprocessed data is given by $f^{sum}_{\mathtt{rma},\mathscr{D}} \circ f^{norm}_{\mathtt{rma},\mathscr{D}}$, and Π^{-1} in Equation (2.9) can be derived by sorting the raw data. To define a `rma` scale adjusting transformation, we therefore need the mean expression values over the sorted arrays ($\hat{\boldsymbol{\mu}}$) as well as the probe specific scales ($\hat{p}_{ik}$) for all probes i and all probesets k.

2.4 Application to data

In this section we assess the benefits of the use of a scale adjusting transformation in signature documentation. We start with describing two measures of a signature's performance and continue by reporting results on eight clinical data sets [10, 12, 14, 65, 107, 118, 127, 153].

2.4.1 Consistency and stability of diagnosis

To contrast the performance of a signature accompanied by a scale adjusting transformation to that of a signature without such a transformation we choose a resampling based approach. The repeated splitting of a data set into a signature deriving part and an external part, to which the signature is applied, mimics real life independent evaluation. Repeating such a process a large number of times brings into play sample variation. In the following, we introduce two performance measures utilizing diagnoses on external patients. *Consistency* focuses on the agreement of diagnoses to a reference, while *stability* compares diagnoses for the same patient derived from signatures inferred from different core data sets.

Consistency We randomly split each of the eight data sets into two parts, which we call the internal and the external set. To ensure comparability of results across studies, the size of the internal sets was fixed to 20 arrays for all data. We then derive a signature using only the internal set and apply it to a random sample of the external set. In case a scale adjusting transformation is present, it is utilized before classifying the external sample. This mimics the process of communicating signatures between health care centers. For evaluation of consistency we determined the diagnosis a patient would have received if analyzed in the context of the original study (reference diagnosis). To this end, we concatenated the external case with the 20 internal arrays, renormalized this complete data set of 21 cases and applied the signature. In the terminology of the last section, let $\mathscr{D}$ denote the internal set. Then we write for a diagnosis with scale adjusting transformation:

$$d^{+} := c_{\mathscr{D}}(f_{\texttt{prep},\mathscr{D}}(A_{n+1})) \quad ,$$

for a diagnosis without scale adjusting transformation:

$$d^{\emptyset} := c_{\mathscr{D}}(\texttt{prep}(A_{n+1}|\emptyset)) \quad ,$$

and for the reference diagnosis

$$d^\star := c_{\mathscr{D}}(\texttt{prep}(A_{n+1}|\mathscr{D})) \quad .$$

We re-run the complete procedure 1000 times using different random partitions of the data into internal and external sets. Assume there are n patients in the study and each is predicted at least m times. For patients predicted more often, a random subsample of size m is taken. Let $\boldsymbol{d} \in \mathbb{R}^{nm}$ be the vector of predictions (either d^+ or $d^{\varnothing}$)and $\boldsymbol{d}^\star$ the reference predictions. Then consistency is defined as

$$\bar{c} = \frac{1}{nm} \sum_i^{nm} \underbrace{\mathbb{I}(\boldsymbol{d}_i = \boldsymbol{d}_i^\star)}_{\boldsymbol{c}_i} \quad . \tag{2.10}$$

Consistency is therefore proportional to the percentage of agreement between a given number of diagnoses and the corresponding reference diagnoses. A consistency of one corresponds to the situation where all diagnoses were identical to the reference. A consistency of zero implies that all diagnoses were different from the reference.

If the entries in $\boldsymbol{c}$, defined in Equation (2.10), arise from iid Bernoulli variables with success-parameter p, then $nm \cdot \bar{c}$ is distributed binomially with parameters p and $N = nm$. Estimates for $\bar{c}$ and confidence intervals follow directly from the binomial distribution. As we are interested in comparing the consistency of signatures with scale adjusting transformation to signatures without scale adjusting transformation, we also look at the quantity $\Delta\bar{c} := \bar{c}_+ - \bar{c}_\varnothing$. The subscripts denote whether a scale adjusting transformation was used. Assuming the same model, $nm \cdot \Delta\bar{c}$ is the difference of two binomially distributed random variables and for the parameters we assume $nm, \bar{c}_+$ and $\bar{c}_\varnothing$, respectively. Performing the convolution numerically yields confidence intervals.

Note that we calculated the confidence intervals assuming independent Bernoulli variables. It is clear that this assumption does not strictly comply with reality: There are dependencies between the diagnoses stemming from the fact that samples are re-used in the internal sets as well as from predicting the same external patient various times.

Correction for class bias In case the classes underlying different diagnoses are not equally frequent, high consistency can be obtained by chance. The κ-coefficient [29] is a measure of agreement which corrects for such chance artifacts. It is defined as

$$\kappa = \frac{P(O) - P(E)}{1 - P(E)} \quad ,$$

where $P(O)$ is the observed agreement (fraction of coinciding diagnoses) and $P(E)$ is the fraction of coinciding diagnoses expected by chance. The expected percentage agreement is calculated from a contingency table by multiplication of the marginal diagnosis-frequencies. This implies assuming independence between the two methods of obtaining diagnoses, e.g. “reference” versus “scale adjusted”. We show the κ-coefficient along with consistency in our results. But as it is non-trivial to interpret [90–92] we focus on consistency in the discussions. In general, consistency and κ-coefficient support the same conclusions we draw from the results we present.

Stability The assessment of stability of diagnoses is based on the same resampling experiment as consistency. But while consistency refers to a comparison of an external diagnosis to a reference, stability compares several external diagnoses for the same patient. For a given patient, we determined the most frequent diagnosis across all resampling-runs. The diagnoses for the same patient may differ, as the corresponding classifiers have been derived from different internal data sets. Let l denote the number of the most frequent diagnosis for a patient. Then $s = 1-(1000-l)/l$ is reported as a stability index. A stability of one corresponds to the situation where the patient was always assigned to the same class. A stability of zero implies that assignments to either class were equally frequent.

More formally, let $\boldsymbol{d}^{(i)}$ be the vector of predictions for patient i. As before, we consider m predictions for each patient. In case there are only two types of predictions let $\boldsymbol{c}_i^0 := \sum_j^m \mathbb{I}(\boldsymbol{d}_j^{(i)} = 0)$ the number of times patient i was predicted to belong to class zero and $\boldsymbol{c}_i^1 := \sum_j^m \mathbb{I}(\boldsymbol{d}_j^{(i)} = 1)$ the number of times patient i was predicted to class one. Then the vector of stability scores $\boldsymbol{s}$ is defined by

$$\boldsymbol{s}_i = 1 - \frac{\min(\boldsymbol{c}_i^0, \boldsymbol{c}_i^1)}{\max(\boldsymbol{c}_i^0, \boldsymbol{c}_i^1)} \quad .$$

If the entries in $\boldsymbol{d}^{(i)}$ arise from m iid Bernoulli variable with parameter p, then the density of $\min(\boldsymbol{c}_i^0, \boldsymbol{c}_i^1)$ is of the form:

$$P(\min(\boldsymbol{c}_i^0, \boldsymbol{c}_i^1) = x) = \begin{cases} \binom{m}{x} p^x (1-p)^{m-x} + \binom{m}{m-x} p^{m-x}(1-p)^x & \text{if } x < m/2 \\ \binom{m}{m/2} p^{m/2}(1-p)^{m/2} & \text{if } x = m/2 \\ 0 & \text{else} \end{cases}$$

and $\max(\boldsymbol{c}_i^0, \boldsymbol{c}_i^1) = m - \min(\boldsymbol{c}_i^0, \boldsymbol{c}_i^1)$. This we use to calculate confidence intervals for the $\boldsymbol{s}_i$. Random stability curves (gray lines in Figure 2.5) were obtained by simulation, assuming $p = 1/2$ for all patients.

2.4.2 Data and results

We studied the impact of a scale adjusting transformation on eight clinical microarray studies [10, 12, 14, 65, 107, 118, 127, 153], involving different disease types and representing diagnostic as well as prognostic classification problems. The issue of specifying a scale adjusting transformation we also call the *signature documentation problem.* The data is summarized in Table 2.1, where we also report on the difficulty of the underlying classification problem and on the gain in consistency when providing a scale adjusting transformation via proper documentation.

We used the two preprocessing schemes vsn and rma we discussed in Sections 2.2 and 2.3, and compared performance with and without $f_{\mathtt{prep},\mathscr{D}}$. For stability assessment we only considered signatures accompanied by a scale adjusting transformation. To distinguish between using and not using $f_{\mathtt{prep},\mathscr{D}}$, we also use the term *documentation by value* and *documentation by reference* . Documentation by value implies supplying quantitative information ($f_{\mathtt{prep},\mathscr{D}}$) that depends on the original signature deriving data set. Documentation by reference refers to the practice of mentioning the preprocessing scheme used for signature derivation, but the lack of additional information. For the vsn preprocessed data we used documentation by reference exclusively, as the model in Equation (2.3) is not identifiable for a single array.

Results on consistency Results comparing the consistency of signatures documented by value to signatures documented by reference are summarized in Table 2.4. We observe in the study by Beer et al. [10] that documentation by reference can lead to discrepancies between external and internal diagnosis being as frequent as 27%. The median consistency across all studies using documentation by reference was 83.5%, corresponding to a median discrepancy of diagnoses as high as 16.5%. This demonstrates the existence and importance of a documentation problem: Diagnoses are unstable and external researchers will generally not obtain the same results as the investigators of the original study. More importantly, we observed that the documentation strategy matters. Documenting signatures by value leads to substantially more consistent results than documentation by reference. We observed the biggest effect for the prognostic study by Beer et al., where consistency improved from 73% to 97%. This corresponds to a consistency gain between 22% and 26% (95% CI). The smallest consistency gain (between 3% and 4%, 95% CI) was observed for the diagnostic study of Willenbrock et al. [153], which poses the most easy classification problem. The median minimal gain in consistency obtained from documenting signatures by value (at 97.5% CI) was 15%. On most data sets consistency of rma and vsn preprocessing were comparable; the differences were small. Exceptions are the data sets of Huang et al. [65] and Ross et al. [118], where consistencies obtained with rma were larger. Overall, signatures documented by value display high consistency, most of them larger than 95%. Documentation by reference was found to be significantly less consistent (about 15% median consistency loss).

Results on stability Results are summarized in Figure 2.5. To the two preprocessing procedures we discussed before (vsn and rma) we added the standard preprocessing of the Affymetrix Microarray Suite (mas). The mas scheme treats arrays independently and does not require a scale adjusting transformation. We assessed the stability of diagnoses (as defined in the previous Section). To display results, we sorted the patients in each study by stability. Values were plotted together with a 75% CI for each patient. For the most difficult classification problems (Beer et al., Bhattacharjee et al. [12], Ross et al.), stability curves increase slowly compared to the curve of the relatively easy diagnostic

Study	Disease	Problem	# Cases	Difficulty	Doc. gain
Beer et al.	Adenocarcinoma	Prognostic	84	difficult	22%
Bhattacharjee et al.	Adenocarcinoma	Prognostic	125	difficult	16%
Huang et al.	Breast cancer	Prognostic	52	medium	14%
Pomeroy et al.	Medullablastoma	Prognostic	60	difficult	16%
Willenbrock et al.	Childhood ALL	Diagnostic	45	easy	3%
Ross et al.	Childhood ALL	Risk Group	87	difficult	16%
Shipp et al.	DLBCL	Prognostic	58	difficult	11%
Bild et al.	Ovarian cancer	Prognostic	133	difficult	12%

Table 2.1: Summary of microarray studies. *Overview of the eight studies used to investigate the signature documentation problem. "Difficult" implies a typical cross-validated success rate of derived signatures of less than 60%, whereas "easy" studies reach a success rate of more than 90%. Documentation gain denotes the increase in consistency when using documentation by value instead of documentation by reference (see text).*

problem from the study by Willenbrock et al.. This indicates a stable diagnosis for almost all patients in the study of Willenbrock et al., whereas in the studies of Beer et al., Bhattacharjee et al., Ross et al., and Bild et al. [14] external diagnoses can vary for a large group of patients. Still, the vast majority of diagnoses are more stable than what is expected by random guessing (gray lines). In the study of Huang et al. the stability curves of `rma` and `vsn` are clearly outperforming `mas`. Note that `vsn` is slightly more stable than `rma`, even though for the same data set consistency was significantly worse. Other studies hinting at differences in stability are Shipp et al. [127], Bild et al. and Pomeroy et al. [107]. All of them see `rma` and `vsn` superior to `mas`. None of the studies provides evidence of the opposite.

Before we discuss our results in Section 2.6, we explore for the `vsn` case whether we can utilize the scale adjusting transformation to assess if comparing an external patient to a core data set is meaningful.

2.5 Compatibility of external patients to core data

Motivation and method The question of whether an external patient is "compatible" to data of a specific core-study has two aspects. Firstly, as stratification of tumor patients is subject of current research and in some cases under discussion [4, 69, 127], there are cases where it is not known a priori if a certain sample fits the context of a study. Secondly, confusion of array labels can lead to improper data being compared to a core study. In both cases it is advantageous to flag such a situation.

Conceptually the question whether an external patient is comparable to a core data set

Dataset		cons.		κ		cons.	κ
		ref	val	ref	val	gain	gain
Beer	rma	$\mathbf{73}^{\uparrow 74}_{\downarrow 71}$	$\mathbf{97}^{\uparrow 97}_{\downarrow 96}$	$\mathbf{45}^{\uparrow 44}_{\downarrow 43}$	$\mathbf{94}^{\uparrow 94}_{\downarrow 93}$	$\mathbf{24}^{\uparrow 26}_{\downarrow 22}$	$\mathbf{49}^{\uparrow 51}_{\downarrow 47}$
	vsn	—	$\mathbf{98}^{\uparrow 98}_{\downarrow 98}$	—	$\mathbf{96}^{\uparrow 97}_{\downarrow 98}$		
Bhattacharjee	rma	$\mathbf{81}^{\uparrow 82}_{\downarrow 79}$	$\mathbf{98}^{\uparrow 99}_{\downarrow 98}$	$\mathbf{62}^{\uparrow 63}_{\downarrow 60}$	$\mathbf{97}^{\uparrow 97}_{\downarrow 96}$	$\mathbf{17}^{\uparrow 19}_{\downarrow 16}$	$\mathbf{35}^{\uparrow 36}_{\downarrow 34}$
	vsn	—	$\mathbf{97}^{\uparrow 97}_{\downarrow 96}$	—	$\mathbf{94}^{\uparrow 94}_{\downarrow 93}$		
Bild	rma	$\mathbf{86}^{\uparrow 87}_{\downarrow 85}$	$\mathbf{98}^{\uparrow 99}_{\downarrow 98}$	$\mathbf{71}^{\uparrow 72}_{\downarrow 70}$	$\mathbf{97}^{\uparrow 97}_{\downarrow 97}$	$\mathbf{13}^{\uparrow 14}_{\downarrow 12}$	$\mathbf{26}^{\uparrow 27}_{\downarrow 25}$
	vsn	—	$\mathbf{98}^{\uparrow 99}_{\downarrow 98}$	—	$\mathbf{97}^{\uparrow 97}_{\downarrow 97}$		
Huang	rma	$\mathbf{87}^{\uparrow 87}_{\downarrow 86}$	$\mathbf{99}^{\uparrow 99}_{\downarrow 99}$	$\mathbf{72}^{\uparrow 73}_{\downarrow 71}$	$\mathbf{98}^{\uparrow 98}_{\downarrow 97}$	$\mathbf{12}^{\uparrow 13}_{\downarrow 11}$	$\mathbf{25}^{\uparrow 27}_{\downarrow 24}$
	vsn	—	$\mathbf{89}^{\uparrow 90}_{\downarrow 89}$	—	$\mathbf{77}^{\uparrow 77}_{\downarrow 76}$		
Pomeroy	rma	$\mathbf{81}^{\uparrow 82}_{\downarrow 89}$	$\mathbf{98}^{\uparrow 99}_{\downarrow 98}$	$\mathbf{61}^{\uparrow 63}_{\downarrow 60}$	$\mathbf{96}^{\uparrow 97}_{\downarrow 96}$	$\mathbf{17}^{\uparrow 19}_{\downarrow 16}$	$\mathbf{35}^{\uparrow 36}_{\downarrow 33}$
	vsn	—	$\mathbf{96}^{\uparrow 97}_{\downarrow 95}$	—	$\mathbf{92}^{\uparrow 93}_{\downarrow 91}$		
Ross	rma	$\mathbf{80}^{\uparrow 82}_{\downarrow 79}$	$\mathbf{98}^{\uparrow 99}_{\downarrow 98}$	$\mathbf{61}^{\uparrow 62}_{\downarrow 59}$	$\mathbf{96}^{\uparrow 97}_{\downarrow 96}$	$\mathbf{18}^{\uparrow 20}_{\downarrow 16}$	$\mathbf{36}^{\uparrow 38}_{\downarrow 34}$
	vsn	—	$\mathbf{92}^{\uparrow 94}_{\downarrow 91}$	—	$\mathbf{85}^{\uparrow 86}_{\downarrow 84}$		
Shipp	rma	$\mathbf{87}^{\uparrow 88}_{\downarrow 86}$	$\mathbf{99}^{\uparrow 99}_{\downarrow 99}$	$\mathbf{73}^{\uparrow 72}_{\downarrow 74}$	$\mathbf{98}^{\uparrow 98}_{\downarrow 98}$	$\mathbf{12}^{\uparrow 13}_{\downarrow 11}$	$\mathbf{25}^{\uparrow 26}_{\downarrow 23}$
	vsn	—	$\mathbf{99}^{\uparrow 99}_{\downarrow 98}$	—	$\mathbf{97}^{\uparrow 97}_{\downarrow 97}$		
Willenbrock	rma	$\mathbf{96}^{\uparrow 97}_{\downarrow 96}$	$\mathbf{00}^{\uparrow 00}_{\downarrow 99}$	$\mathbf{92}^{\uparrow 93}_{\downarrow 92}$	$\mathbf{99}^{\uparrow 99}_{\downarrow 99}$	$\mathbf{3}^{\uparrow 4}_{\downarrow 3}$	$\mathbf{7}^{\uparrow 7}_{\downarrow 6}$
	vsn	—	$\mathbf{00}^{\uparrow 00}_{\downarrow 99}$	—	$\mathbf{99}^{\uparrow 99}_{\downarrow 99}$		

Figure 2.4: Documentation by value increases consistency. *Each row contains results on consistency and the κ-coefficient for one of the eight clinical microarray studies. For the preprocessing schemes `rma` and `vsn`, we report consistency indices and κ-coefficients for signatures documented by reference (left columns) and signatures documented by value (right columns), respectively. The last two columns show the improvement achieved through documentation by value. The sub-and superscripts denote 95% confidence intervals. Documentation by value significantly increases consistency and κ-coefficient in all studies.*

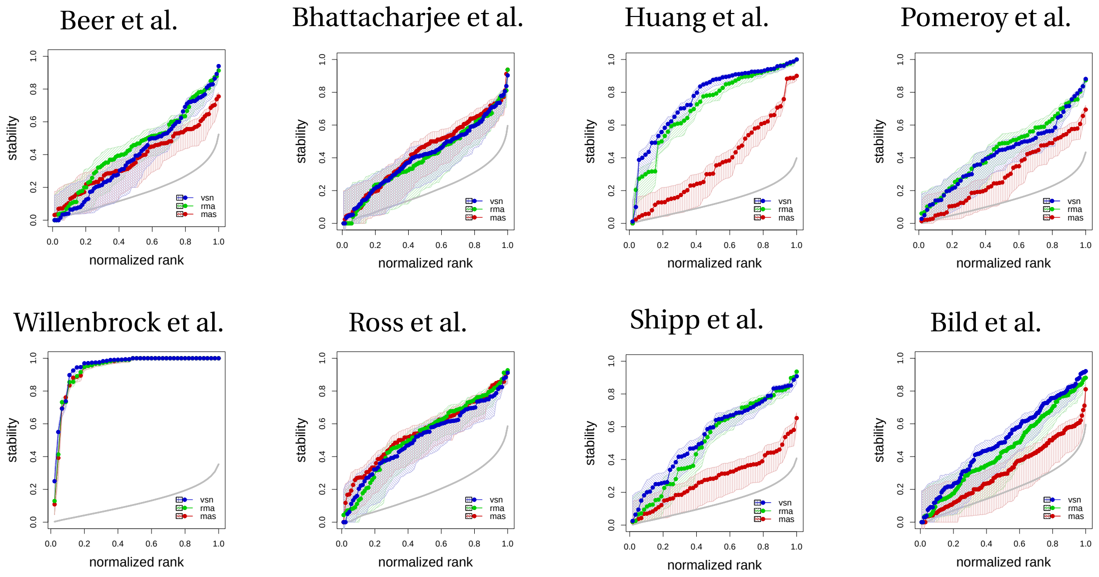

__Figure 2.5: The stability of signatures depends on the preprocessing scheme__. Each plot shows the stability of fully documented signatures in a clinical microarray study. For three preprocessing schemes: `mas` (red), `rma` (blue) and `vsn` (green), we show stability indices on the y-axis. All indices are sorted, such that the x-axis holds ranks, and the curves are guaranteed to increase. The shaded areas show 75% confidence intervals while the gray line corresponds to expected stability indices obtained from random guessing. Stability depends on the preprocessing scheme, with `rma` and `vsn` outperforming `mas` (see text).

can be viewed as an *outlier detection* problem: Are the core data set and the new patient samples of the same distribution (disease population)? Outlier detection problems also arise in many other applied settings including network intrusion, fraud detection, fault detection (in manufacturing processes), marketing and customer segmentation [7]. It is an active field of research [7, 120] with recent methodological advances [2, 9, 86].

We briefly explore the possibility of utilizing the probabilistic model of Equation (2.7) to decide whether comparing an external patient to a `vsn` normalized data set seems reasonable. While this might be suboptimal compared to utilizing one of the generic methods available for outlier detection, its is appealing because it neatly integrates into the `vsn` preprocessing framework and comes with no additional computational cost.

Motivating the scale adjusting transformation for `vsn` preprocessed data we introduced the following model (see Equation (2.7)) for expression values $x_{k(n+1)}$ of an external patient:

$$x_{k(n+1)} = h_{n+1}(x^{\texttt{raw}}_{k(n+1)}) = \hat{\mu}_k + \epsilon_k, \quad \epsilon_k \sim N(0, \hat{\sigma}^2_\epsilon) \;\text{ for }\; k \in \mathcal{K}.$$

Here we have used the notation from Section 2.2 and consider the data only up to probe level. We now ask if the likelihood (corresponding to the above model, see Equation (2.8)) of the new patient data indicates whether this patient might be compared to the core data set in a meaningful way. To explore this empirically we compiled a list of eleven cancer data sets [8, 10, 14, 25, 26, 56, 65, 104, 125, 131, 157], all utilizing (at least partially) the same Affymetrix GeneChip (version HGU95AV2). From each data set we then randomly chose a core data set and `vsn` normalized it to set the scale. Subsequently we transformed each of the remaining patients to this scale, keeping track of the likelihood.

Results and conclusion The results are displayed in Figure 2.6. From the datasets with the gray background (a breastcancer and two leukemia data sets) 40 samples (30 samples in the lowermost plot) were taken as core data. All remaining patients from all data sets were `vsn` background-corrected and normalized to fit the core data. For each core data set we decided on a cutoff for the likelihood; patients with sufficiently high likelihood are considered similar. Our criterion for the cutoff choice was the correct assignment of roughly 95% of the external patients taken from the same data set as the core data. The percentage of similar patients for each data set is is marked on top of the plots. With a carcinoma core data set (the breast cancer data of Huang et al. [65]), the log-likelihood separates patients with carcinoma from patients with leukemia (uppermost plot). Also, the only other breast cancer data set [14] is the one with the most similar patients. The converse does not really hold, as we can see from the lower plots. In the middle plot the majority patients from all data sets, except from the mixed-lineage leukemia (MLL) data set of Armstrong et al. [8], appear similar to the leukemia data from Cheok et al. [25]. For the lowermost plot we used the MLL data as core data set; surprisingly two carcinoma data sets [10, 131] and one glioma data set [104] appear most similar. Overall, the experiment suggests that a low likelihood *can* be a good rejection criterion, but this does not necessarily have to be the case.

2.6 Discussion and chapter summary

To the best of our knowledge, the problem of documenting diagnostic expression signatures has not been pointed out and studied before. The reason might be that analyses are often conducted in a homogeneous study environment and do not face the documentation problem. Documentation of a signature comes fully into play only by independent evaluation of a given classification rule. We were able to demonstrate that common documentation standards are insufficient for unambiguously determining diagnosis. We observe low average consistency values for the documentation by reference strategy. We have shown that the consistency of diagnostic signatures can be improved substantially by documenting data-dependent preprocessing information. To do so, we have proposed the documentation by value strategy providing a scale adjusting transformation in addition to the signature. We observed a trade-off between the performance of preprocessing protocols as reported in [30, 75] and the effort required

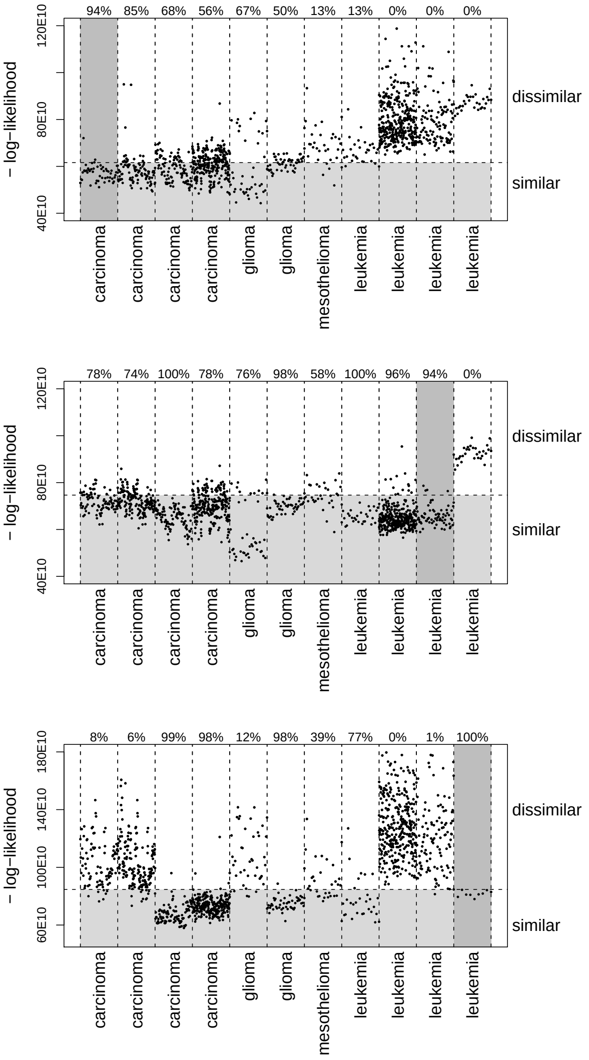

Figure 2.6: The log-likelihood as a compatibility criterion. *In the plots on the left we show the negative log-likelihood (Equation (2.8)) for external patients of eleven different cancer data sets ([8, 10, 14, 25, 26, 56, 65, 104, 125, 131, 157], from left to right). From the datasets with the gray background 40 samples (30 samples in the lowermost plot) were taken as core data; all patients were background corrected and normalized to this scale. For each plot we decided on a similarity threshold based on high sensitivity (see text). The percentage of patients similar to the core data is marked on top. With a carcinoma core data set the log-likelihood separates carcinoma from leukemia (uppermost plot). The converse does not really hold, as we can see from the lower plots on the left. Overall, the experiment suggests that a low likelihood* can *be a good rejection criterion, but this does not necessarily have to be the case.*

for documenting them. While it is known that preprocessing schemes sharing information across arrays can enhance precision and accuracy of estimated expression differences [30, 75], improved normalization performance comes at a price: The already normalized expression values for a fixed microarray change when additional arrays are added to the study (see Section 2.3). This is a problem for applying a signature to external data; the original data needs to be included in the normalization of external arrays. The re-normalization of the complete data set changes the original expression values, affecting the signature and the molecular diagnosis of patients in the original study.

To circumvent this problem, we have altered the widely used preprocessing methods `vsn` [68] and `rma` [72] to provide an "add-on" mode. This mode allows to process a core data set deriving a scale adjusting transformation. This transformation can then be utilized to add data from additional arrays without changing the normalized core data.

As a summary of our findings we propose the following (general) guidelines for deriving and *documenting* a diagnostic gene expression signature:

1. *Preprocessing* Preprocess the data using a protocol that allows for later inclusion of arrays without changing the original expression values. For example, preprocess arrays independently of each other, or by providing a scale adjusting transformation.

2. *Building the classification rule* Derive a classification rule using software that provides a complete quantitative specification of the signature for documentation purposes. For example, use the nearest shrunken centroid procedure [139] we employed.

3. *Documentation by value* Document the full quantitative specification of the classification rule. In addition, document preprocessing. For example, use the software we provide and document it in form of a scale adjusting transformation. Ideally, publish both parts as an integrated open source computer program that can readily be used to diagnose new patients.

4. *Diagnosing an external patient* Bring the raw data to a signature consistent scale. Apply the documented classification rule to diagnose the new patient.

These guidelines suggest methods we have found to work well in practice, but we do not claim them to be optimal in any sense. Given the heterogeneity of clinical data as well as the diversity of array platforms, it can safely be assumed that there is data where other methods are more appropriate. However, we believe that in these situations the documentation problem still exists, and a similar documentation by value strategy should be developed for the methodology in use.

In our simulation setup, the data we call external are actually arrays from the same study. With real external data, additional problems occur. It has been shown that even when using the same technology and experimental protocols, the resulting data for the same tissue sample varies between different health care centers [74]. While this effect is not directly linked to documentation, we believe that the benefits of documenting signatures by value are enhanced in situations where external and internal data are more heterogeneous. Documentation of signatures is significantly easier for preprocessing methods treating arrays independently of each other, as is the case for the Affymetrix Microarray Suite (`mas`). However, we do not recommend these methods due to inferior normalization performance and the reduced stability of signatures we observed.

While microarray based diagnostic signatures hold great promise to improve diagnosis and prognosis of disease, evaluation of a signature's predictive performance is difficult and subject to much current research and argument [13, 99, 114, 141, 150]. It is important to prove that a signature holds independent complementary information to existing prognostic markers. No gene expression signature has reached this status [41, 130]. While sharing candidate signatures within the research community can accelerate the process of evaluation, this does not allow for any ambiguity of signatures. We believe that our documentation by value strategy removes this obstacle and greatly facilitates this endeavor. The additional effort required is small. There are certainly several ways to implement documentation by value. We have shown one of them.